Computational Materials Science 134 (2017) 190–200

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

Error estimation in high-throughput density functional theory calculation for material property: elastic constants of cubic binary alloy case

Juan Wang [a,b], Xiaoyu Yang [a,b,*], Guisheng Wang [c], Jie Ren [a,b], Zongguo Wang [a], Xushan Zhao [a], Yue Pan [d]

[a] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China
[c] College of Applied Sciences, Beijing University of Technology, Beijing 100124, China
[d] School of Mathematics, Sun Yat-Sen University, Guangzhou, 510275, China

ARTICLE INFO

Article history:
Received 23 July 2016
Received in revised form 14 December 2016
Accepted 22 March 2017

Keywords:
High-throughput DFT calculation
Error estimation
Neural network
Cross validation
Support vector regression

ABSTRACT

Estimating Density Functional Theory (DFT) calculation error is an important while challenging task in computational material science. The calculation contains inherent errors due to improper input parameters and approximated exchange-correlation functional. In this paper, we present a data-driven approach of using machine learning techniques to estimate the error of DFT calculation. We prepare the data by high-throughput first principle DFT simulation and experimental data collection. The single-hidden layer back propagation feedforward neural network (SLBPFN) constructed based on the proposed cross validation algorithm, and support vector machine for regression (SVR) techniques are employed to build regression models to predict the DFT calculation error. As a demonstration, the developed regression models are used to predict errors in calculating elastic constants of cubic binary alloys. It has been demonstrated that the machine learning techniques can predict DFT calculation error of elastic constants with an acceptable accuracy. It also shows the BP neural network built by our proposed cross validation algorithm can provide a better prediction. Our study is a first-invasive work of using machine learning techniques to estimate the errors in calculating elastic constants of binary alloys.

1. Introduction

With the wide applications of DFT [1] in prediction of materials properties, estimating error of the DFT calculation result (DFT error) becomes increasingly urgent. The DFT calculation results can provide references to plan future experiments or to find materials with desired properties. The databases currently being created by high-throughput frameworks, such as AFLOW [2,3], Materials Project [4,5] and MatCloud [6] are used for further intelligent data mining. The automatic flow computation of high-throughput makes the materials calculations easier and faster from ideas to results. Using the DFT simulation data combined with machine learning techniques, many research achievements have been achieved in the prediction of material properties [7–9]. With the material simulation data being widely used, the error of simulation data becomes a concern. It is therefore essential to find an effective method to assess the deviation between an individual DFT prediction of a certain property and the corresponding experimental measurement. Error estimations in DFT calculation are usually provided by statistical analysis methods [10,11], where the statistics are used to represent the statistical error of the calculated materials properties, such as standard deviation and confidence interval.

For an *ab initio* simulation package, the input parameters are the main error sources for a prediction, and proper parameters setting produces smaller error. Material property calculations using DFT software packages integrated in high-throughput frameworks depend on a large number of parameters, such as exchange-correlation (XC) functional, pseudopotential, *k*-point grid density, plan wave kinetic energy cut-off, convergence accuracy, and so on. The exact XC functional is unknown [12], and all DFT calculations adopt approximated XC functional, which lead to inevitable DFT error. Because of computational cost, finite basis sets are chosen in practical DFT calculations, the inadequate basis sets are not able to cover entire physical space and this inadequacy also introduces inherent computational error. The parameters *k*-point, cut-off and convergence accuracy can also introduce numerical DFT

* Corresponding author at: Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China.
E-mail address: kxy@cnic.cn (X. Yang).

http://dx.doi.org/10.1016/j.commatsci.2017.03.035
0927-0256/

error. In DFT, calculations using different parameter values for *k*-point, energy cut-off, XC functional and pseudopotential *etc.*, yield unidentical DFT errors. Due to the intrinsic complexity of DFT codes it is impractical to carry out DFT calculations with the full set of calculation parameters, and therefore it is also impractical to investigate DFT error caused by each set of parameters through DFT calculation. Finding a method to estimate the DFT calculation error caused by each set of parameters is meaningful, which can reduce the total DFT calculation time for finding proper parameters. The set of parameters that produces minimum DFT error is recommended to calculate material property in action.

Machine learning techniques and DFT simulation data of elastic constants for materials have been applied to predict elastic properties of compounds or to design materials with maximal or minimal elastic properties. Artificial neural network is employed to build a prediction model of the mechanical properties of Cu–Sn–Pb–Zn–Ni cast alloys using chemical compositions [13]. Gaussian Process Model (GPM) and SVR are used to build models that predict elastic properties in terms of elementary orbital radii of the individual components of compound materials [14], and the initial training data they used are calculated by DFT [15]. It can be seen that elastic constants predicted by DFT calculations are increasingly used by material researchers, therefore it is intelligible to determine justifiable error estimations for elastic constants to obtain lower error DFT predictions.

In this paper, we proposed to use a data-driven machine learning approach to estimate the DFT calculation error of elastic constant of cubic binary alloy. Why chose elastic constant? In general, some material properties such as elastic constant are closely related to the crystal structure, while some material properties such as band structure and dielectric constant are closely related to electronic-level information. In comparison to crystal information, microscopic electronic-level information is usually hard to obtain. Hence we chose elastic constant calculation in our case as an entry.

The approach is based on the following assumption. As a compound is composed of pure elements, we think that elastic constant calculation error of a compound could be deduced from the elastic constant calculation errors of its comprising pure elements (using same calculation parameters), and the crystal structure similarities between this compound and its comprising pure elements. We understand that for different materials, the systematic DFT errors with different exchange-correlation functions are often significantly different [16]. As the number of pure elements is limited, the elastic constant calculation error of each pure element under different calculation parameters can be easily obtained as follows: (i) using high-throughput computing facilities (e.g. MatCloud) to calculate the elastic constant under different combination of parameters; (ii) collecting experiment value of elastic constant from literature; (iii) working out the elastic constant calculation error by comparing the calculation value and experiment value.

The procedures of the proposed approach are described as follows: (1) We collect experimental data of elastic constants for 45 binary alloys with cubic crystal system and 63 elemental metals from literatures, which are used to compute the DFT errors for corresponding material, and then use MatCloud to calculate physical properties for pure elements and binary alloys, which are used as the predictors of the machine learning models; (2) We briefly introduce two kinds of machine learning techniques, SLBPFN and SVR, which are used to build regression models predicting error of DFT calculation elastic constants caused by improper parameters and functional; (3) Finally, based on the collected data set, the SLBPFN and SVR techniques are used to construct prediction models to estimate DFT errors of elastic constants for cubic binary alloys without DFT code runs, and results by the two prediction models are compared.

The remainder of this paper is organized as follows. In the Section entitled Methods, we describe the data preparation for using machine learning technique, and discuss two machine learning methods as well as data set for building DFT error prediction model. The next section focuses on the analysis of regression and application performance of the proposed approach. We conclude in Section 4.

2. Methods

2.1. Data preparation

Data preparation is a crucial stage in using machine learning. The source and purpose of the collected data set are listed in Table 1. The predictions of DFT error of elastic constants for binary alloys are built from a data set consists of experimental elastic constants of 45 binary alloys with cubic crystal system and 63 pure elements, and predictors for the 45 binary alloys. The predictors include elemental information, structure information and simple physical properties of pure elements and binary alloy, and DFT errors of elastic constants for associated pure elements. The predictors of simple physical properties, such as volume and energy are obtained from DFT calculations.

In the present work, DFT calculations are performed by the VASP code [17,18]. The ion-electron interaction is described by the projector-augmented wave (PAW) method. The used XC functionals are depicted by the generalized gradient approximation (GGA) parameterized by Perdew and Wang (PW91), Perdew Burke Ernzerh (PBE), and Local Density Approximation (LDA). The sampling values for VASP PREC tag are *normal*, *high* and *accurate*. The PREC tag determines the energy cut-off, if (and only if) no value is given for energy cut-off in calculation parameters file. For PREC = *normal* and PREC = *accurate*, energy cut-off will be set to the maximal energy cut-off value found in the functional file. For PREC = *high*, energy cut-off is set to the maximal energy cut-off value in the functional file plus 30%. We did a convergence test, and found that when the total number of k-points (KPT) in the irreducible Brillouin zone is set to 21,000, the DFT calculation results

Table 1
The data source and purpose of the collected data.

Data to be prepared	Data source	Purpose
Volume of 63 pure elements	By calculation	Used as predictors of machine learning model
Energy of 63 pure elements	By calculation	Used as predictors of machine learning model
Elastic constant of 63 pure elements	By calculation and collecting experimental data	Used to work out the error of elastic constant calculation, which is used as predictors of machine learning model
Energy of 45 binary alloys	By calculation	Used as predictors of machine learning model
Elastic constant of 45 binary alloys	By calculation and collecting experimental data	Experimental data is used to calculate error of calculated elastic constants. Error of calculated elastic constant is used as target variable of machine learning model